BIOINFORMATICS ORIGINAL PAPER
Vol. 22 no. 23 2006, pages 2890–2897
doi:10.1093/bioinformatics/btl492

Gene expression

A scalable method for integration and functional analysis of multiple microarray datasets

Curtis Huttenhower, Matt Hibbs, Chad Myers and Olga G. Troyanskaya*

Department of Computer Science, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA

Received on July 31, 2006; revised on September 5, 2006; accepted on September 19, 2006
Advance Access publication September 27, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: The diverse microarray datasets that have become available over the past several years represent a rich opportunity and challenge for biological data mining. Many supervised and unsupervised methods have been developed for the analysis of individual microarray datasets. However, integrated analysis of multiple datasets can provide a broader insight into genetic regulation of specific biological pathways under a variety of conditions.

Results: To aid in the analysis of such large compendia of microarray experiments, we present Microarray Experiment Functional Integration Technology (MEFIT), a scalable Bayesian framework for predicting functional relationships from integrated microarray datasets. Furthermore, MEFIT predicts these functional relationships within the context of specific biological processes. All results are provided in the context of one or more specific biological functions, which can be provided by a biologist or drawn automatically from catalogs such as the Gene Ontology (GO). Using MEFIT, we integrated 40 *Saccharomyces cerevisiae* microarray datasets spanning 712 unique conditions. In tests based on 110 biological functions drawn from the GO biological process ontology, MEFIT provided a 5% or greater performance increase for 54 functions, with a 5% or more decrease in performance in only two functions.

Contact: ogt@cs.princeton.edu

Supplementary information: Supplementary data, a collection of predictions made by MEFIT and software implementing MEFIT are available online at http://function.princeton.edu/mefit/.

1 INTRODUCTION

Within the past decade, biological datasets have become available spanning not just whole genomes but multiple genomes, both within and across species. In particular, microarray coexpression studies routinely profile whole genomes simultaneously; and with shrinking costs, thousands of whole-genome experiments have become publicly accessible for many model organisms. Many methods have been proposed for extracting biological meaning from microarray data, including normalization and meta-analysis (Choi *et al*., 2003; Griffith *et al*., 2005; Hu *et al*., 2005; Moreau *et al*., 2003), clustering (Allison *et al*., 2006; Butte *et al*., 2000; Cheng and Church 2000; Eisen *et al*., 1998; Heyer *et al*., 1999), signature algorithms (Bergmann *et al*., 2003; Ihmels *et al*., 2005; Ihmels *et al*., 2002; Kloster *et al*., 2005), detection of differential expression (Baggerly *et al*., 2001; Cui and Churchill, 2003; Ideker *et al*., 2000), and many others. Although complete analysis of individual microarray datasets is by no means a solved problem, it is of interest to begin examining the additional conclusions derivable from analysis of many microarray datasets. Integration such as this can enable broader understanding of gene regulation in the context of specific pathways and can allow the discovery of coexpression relationships too weak to be detected in individual experiments.

Such integrated analysis of microarray datasets is challenging because of differences in technology, protocols and experimental conditions across datasets. Thus, any microarray integration system must be robust to such differences and should easily adjust to new datasets, perhaps from technologies yet to be developed. Furthermore, in examining any diverse biological datasets (such as microarray results drawn from differing experimental conditions), it is critical to consider functional specificity, i.e. which biological processes are active in which experiments (Huttenhower and Troyanskaya, 2006). For example, in a set of a thousand microarray experiments over *Saccharomyces cerevisiae*, only 10 experiments might have been performed under conditions inducing sporulation. These few microarrays might show strong coexpression of meiotic genes not expressed or not coregulated under other circumstances. This is a benefit in that it provides more specific information regarding meiosis-related genes, but such a relatively small signal can easily be lost during data processing. The problem of integrating many high-throughput data sources thus includes a problem of determining functional relevance; not only can such data reveal genes that are functionally related, it can also reveal the biological circumstances under which they relate.

To this end, we propose a Microarray Experiment Functional Integration Technology, MEFIT; this is a Bayesian framework facilitating the integration of multiple microarray datasets for predicting coexpression-based functional networks of proteins. Furthermore, each of MEFITs predicted functional relationships is provided within the context of a specific biological process. These biological functions of interest can be provided directly by a biologist, or they can be derived automatically from functional catalogs such as the Gene Ontology (GO) (Ashburner *et al*., 2000) or MIPS (Ruepp *et al*., 2004). In addition to its predicted functional relationships, MEFIT's analysis process also provides a functional association score indicating how predictive each input microarray dataset is of each biological function.

*To whom correspondence should be addressed.

Most prior work in large-scale microarray integration has been performed in one of two contexts: statistical meta-analysis or the introduction of multiple microarray experiments into heterogeneous data integration systems. Choi *et al*., 2003; Rhodes *et al*., 2004; Hu *et al*., 2005 and Mulligan *et al*., 2006 are representative examples of the former, all of which use meta-analysis to integrate microarray experiments for the detection of differential gene expression. In MEFIT, we take advantage of similar meta-analytic techniques in order to make disparate microarray experiments comparable, but we build upon the results to make predictions of global coexpression relationships and biological function and to determine the functional specificity of input microarray datasets.

Pavlidis *et al*., 2002; Clare and King, 2003; Troyanskaya *et al*., 2003; Lee *et al*., 2004 and Butte and Kohane, 2006 describe methods for the use of heterogeneous data integration to predict gene function or functional relationships, but none of these (or similar) systems focus specifically on the way in which microarray experiments are integrated. Most often, correlation over individual datasets or all datasets simultaneously is used with minimal inter-study normalization. This can result in a surprising amount of lost information, particularly since microarrays often represent by far the most extensive body of data available for integration (Karaoz *et al*., 2004; Lee *et al*., 2004; Pavlidis *et al*., 2002; Troyanskaya *et al*., 2003). MEFIT improves on these prior systems by providing a scalable integration system specifically for microarrays that takes advantage of the functional diversity of coexpression data to improve prediction accuracy, to provide additional biological context for predicted functional relationships, and to identify biological functions in which individual datasets are particularly informative. To our knowledge, none of these prior systems has provided a means of predicting both gene pair functional relationships and the specific biological processes in which those interactions are expected to occur.

The MEFIT system predicts functional relationships between genes within individual biological processes, consuming microarray datasets and known functional annotations as input. These predictions are generated as probabilities using a Bayesian framework trained in a function-specific manner. This training process allows one to derive relevance scores from the learned network parameters indicating how reliable the system finds each dataset to be within each biological process. Thus, MEFIT determines which microarray conditions are informative for a particular biological function in addition to predicting process-specific functional relationships.

2 METHODS

The primary outputs of the MEFIT system are predicted probabilities of gene pair functional relationships within individual biological functions. These coexpression networks are derived from naive Bayesian networks trained on a per-function basis using microarray data and known functional annotations. The learned parameters of these networks also contain information regarding how predictive each microarray dataset is of each biological function. Biological functions of interest are provided to the system as simple gene sets (i.e. lists of genes annotated to processes, such as *mitotic cell cycle*, or pathways, such as *fatty acid biosynthesis*), which are used to generate known positive pairwise relationships. Known unrelated gene pairs (negatives) are provided as a separate input to the system. For these experiments, we use functional annotations from the GO (Ashburner *et al*., 2000) to generate both positive and negative gene pairs.

As Figure 1 summarizes, microarray data are preprocessed in order to serve as observations during training and evaluation of MEFIT's collection of naive Bayesian networks. The input functional annotations are used to derive known functional relationships (for training and evaluation) as well as to provide functional specificity by dictating the biological contexts for which separate Bayesian networks should be constructed. Finally, naive Bayesian inference on these per-function networks serves to produce predicted functional relationships for gene pairs within each biological function.

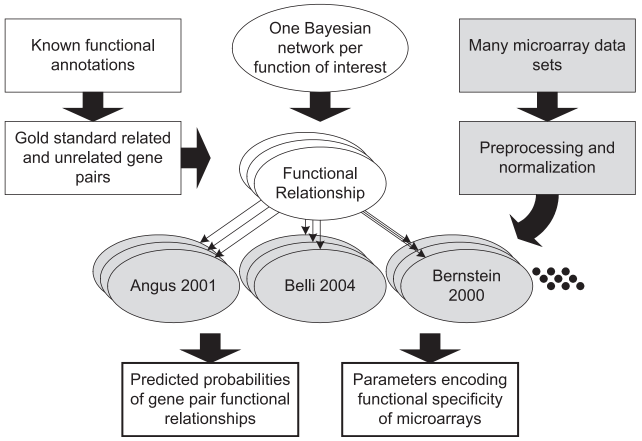
Known functional annotations
One Bayesian network per function of interest
Many microarray data sets
Gold standard related and unrelated gene pairs
Functional Relationship
Preprocessing and normalization
Angus 2001
Belli 2004
Bernstein 2000
Predicted probabilities of gene pair functional relationships
Parameters encoding functional specificity of microarrays

Fig. 1. An overview of MEFIT. A schematic of MEFITs data processing and control methodology. Microarray datasets are provided as input; these are preprocessed and quantized to serve as inputs for naive Bayesian networks. A single network structure is used for all biological functions, but the parameters of these networks are trained individually for each function of interest. Biological functions of interest and known FNs for training are derived from input sets of functional annotations. Finally, Bayesian inference produces a probability of functional relationship for each gene pair within each biological function.

2.1 Microarray data preparation

We assembled *S.cerevisiae* microarray data available from the Stanford Microarray Database (Ball *et al*., 2005), the NCBI Gene Expression Omnibus (Barrett *et al*., 2005), and several independent sources (see Supplementary Table 1 for a complete list). These data comprised 40 unique datasets (time courses or other cohesive collections of experiments) drawn from 34 publications for a total of 712 individual experiments, some single and some dual channel. For each individual dataset, genes missing in >50% of the experimental conditions were removed, and the remaining missing values were imputed using KNNImpute (Troyanskaya *et al*., 2001) with $k = 10$. Finally, replicated genes were averaged to ensure that each dataset contained at most one expression vector per open reading frame.

For single channel data, expression values <2 were considered to be missing, and all single channel values were logarithmically transformed as a final preprocessing step. Since mismatch hybridization values were not available in many datasets, they were not used in this analysis.

Within each dataset, we calculated Pearson correlations between every pair of genes. These correlations were then normalized using Fisher's Z-transform (David, 1949):

$$Z = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right)$$

This maps a correlation ρ into a Z-score, where the collection of pairwise Z-scores within a dataset is guaranteed to be normally distributed. After dividing by the dataset standard deviation and subtracting the mean, this distribution will be $N(0, 1)$, making cross-dataset analyses more robust. Thus, from each microarray dataset, we produce a collection of gene pair Z-scores representing the number of standard deviations their correlation lies away from the mean.

2.2 Bayesian data integration

MEFIT integrates multiple microarray datasets in specific biological contexts to allow for greater accuracy when predicting functional relationships. For each biological function of interest, MEFIT uses a naive Bayesian model to combine many microarray datasets and produce a single, integrated functional relationship score for every pair of genes, creating a function-specific, probabilistic coexpression network. Thus, a separate Bayesian network is trained for each process or function of interest. Each of these networks generates predicted functional relationships within its particular biological function, and the parameters of the network encode how informative a dataset is within that function; an individual dataset is likely to provide varying degrees of predictive accuracy across disparate biological functions.

In each network, the probability of each dataset's observed correlation (represented as *Z*-scores) is conditioned on the probability of functional relationship; each dataset's *Z*-scores are discretized into five bins (below -1, -1 to 0, 0 to 1, 1 to 2 and above 2) and assigned to a single node in the model. Finer binning was found to lead to overfitting and data sparsity issues (data not shown). This results in a Bayesian network with one node (*FR*) predicting functional relationships and *n* nodes conditioned on *FR*, each representing the value of some dataset D_i. For some gene pair (g_i, g_j) and supporting data $\{d_1(g_i, g_j), d_2(g_i, g_j), \ldots, d_n(g_i, g_j)\}$ with $d_k(g_i, g_j) \in \{0, 1, 2, 3, 4\}$, the probability of functional relationship is thus:

$$P_{i,j}\,(\mathrm{FR} = \mathrm{yes}) \propto \prod_{k=1}^{n} P\left[D_k = d_k\left(g_i, g_j\right)\right]$$

The University of Pittsburgh Decision System Laboratory's SMILE library and GENIE modeling environment (Druzdzel, 1999) employing the Lauritzen inference algorithm (Lauritzen and Spiegelhalter, 1988) were used for Bayesian network manipulation, in addition to our own C++ implementations of basic naive Bayesian learning and inference (Neapolitan, 2004). After training (discussed below), this method produces one probability of functional relationship per gene pair, referred to in this paper as the BAYESIAN integration process.

2.3 Testing and validation

MEFIT requires one or more sets of functionally related genes as input (positives), as well as a collection of unrelated gene pairs (negatives). For these experiments, we used 200 functions drawn from GO as positive sets (Supplementary Table 2); genes coannotated below these terms were considered to be functionally related. These terms were hand selected by a panel of six yeast genetics experts who were asked to evaluate whether each GO term would be informative enough to direct laboratory experiments. We constructed a set of all GO terms receiving four or more votes and added their descendants; we then trimmed this set by discarding any term for which all paths to the ontology root were blocked by another term in the set. Any pair of genes sharing an annotation beneath some term in this set was considered to be related (positive). To generate negative examples, any gene pairs not coannotated below some GO term including at least 10% of the *S.cerevisiae* genome (roughly 645 genes) were considered to be unrelated (Supplementary Table 3). This resulted in a set of 619278 related and 8853875 unrelated pairs.

Of the genes included in this set of pairs, 20% (951 genes) were randomly selected as test genes and held out of all training. Our test set consisted of any pair including at least one of these genes (241408 positive and 3320786 negative pairs), and the remaining pairs were used for training. Performance was evaluated using areas under the ROC curves (AUC). All AUCs were calculated analytically using the Wilcoxon Rank Sum formula (Lehmann, 1975). We generally observed only small differences between training and test performance (Supplementary Table 4). These training and test sets were used for the construction and validation of the global integration methods discussed below, and they were further subdivided for per-function analyses.

2.4 Global microarray integration

As discussed above, we implemented a version of our system that trains only one global network, referred to as BAYESIAN integration. For comparison purposes, we also implemented three non-Bayesian integration methods. Most naively, after preprocessing up to the gene averaging stage (excluding Fisher's Z-transform), each microarray dataset was individually normalized per gene to have mean zero and standard deviation one. After this, all experiments were concatenated to create one large expression vector per gene, and pairwise Pearson correlations were calculated using these vectors. For each gene pair, conditions in which at least one missing value remained (due to genes not present in particular datasets) were removed from the correlation calculation. This resulted in the CONCATENATION integration technique.

One can also integrate microarray datasets using statistical meta-analysis similar to that discussed in (Choi *et al*., 2003). To accomplish this, we proceeded through preprocessing as described above to the point where each dataset was represented by a collection of pairwise *Z*-scores drawn from *N*(0, 1). For each gene pair, these *Z*-scores were averaged over the datasets including that pair, producing the Z-SCORE integration data.

Finally, we implemented a version of the microarray integration method discussed in (Lee *et al*., 2004) in the context of general data integration. In brief, pairwise Pearson correlations were calculated per dataset. The pairs in the training set were used to produce a modified precision/recall plot (a log-likelihood (LLS)/correlation plot) to which a sigmoid curve could be fit. This curve allowed transformation of correlations from the test set into a LLS space from which datasets are integrated by taking the average LLS for a gene pair across all available data. This will be referred to as the LLS integration.

2.5 Functional analysis

For microarray integration on a per-function basis, it was necessary to further decompose the training and test sets into collections of gene pairs relevant to each biological process of interest. In all cases, a gene pair was considered relevant to some function if (i) it represented a positive relationship and both genes were included in the function or (ii) it represented a negative relationship and one gene was included in the function. This provides a definition of training and test sets for each function provided as input (e.g. the GO terms discussed above).When evaluating the performance of the four global integration techniques on individual functions, training (for the BAYESIAN and LLS methods) was performed using the entire training set. Evaluation was performed using each function's test set. Functions containing fewer than ten genes (45 gene pairs) were not considered during testing.

Using the same naive Bayesian framework, we also learned one network per-function using the individual training sets—the MEFIT integration technique. In addition to the predictive benefits discussed below, this provided additional information relating each microarray dataset to each function of interest. Specifically, we calculated the average difference between the prior and posterior probabilities of a functional relationship for each dataset and function. For each biological function, dataset and discretized value, we provided only that datum as input to the function's Bayesian network. We then averaged (over the five discretized inputs for a particular dataset) the absolute values of the differences between the networks's prior and the posterior probabilities of functional relationship generated in this manner. This provides a measure of how 'trustworthy' or influential each dataset is when predicting gene pairs in each function (Fig. 4).

3 RESULTS

3.1 Characteristics of functional relationships vary by biological process

The performance of the five integration techniques on selected GO functions can be seen in Figure 2, and Supplementary

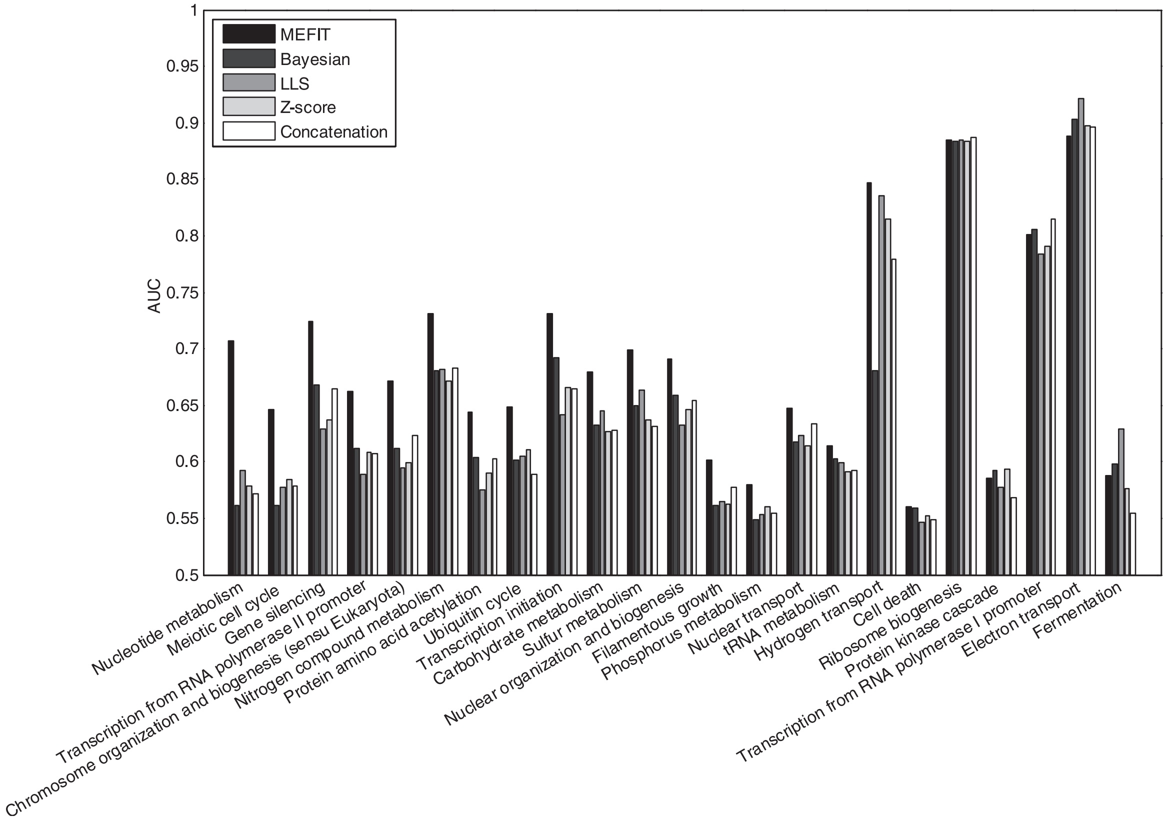

Fig. 2. Per-function performance of each integration technique. Areas under sensitivity/specificity curves (AUCs) for a selection of biological functions extracted from GO, ordered from most to least improvement and evenly spanning MEFIT's performance range. MEFIT showed an AUC increase of 5% or more over all other integration techniques in 54 of the 110 functions evaluated; AUC decreased by 5% or more in only two functions. AUC values range from random at 0.5 to optimal at 1.0. We measured performance for the concatenation, *Z*-score, LLS, Bayesian and MEFIT integration techniques; results for all GO terms in our answer set appear in Supplementary Table 4.

Table 4 contains a complete listing. The MEFIT integration method yields an AUC increase of 5% or more (over the maximum of the other four methods) in 54 of the 110 functions for which evaluation of all five methods was possible. Performance increased by a smaller amount in 31 of the remaining functions and decreased by >5% in only two functions.

Interestingly, the functions in which the simpler CONCATENATION and Z-SCORE techniques perform well relative to the other three integration methods are also among those with the highest overall AUCs: *ribosome biogenesis*, *rRNA metabolism*, *RNA methylation*, *electron transport* and *cellular respiration*. This may indicate that for such high-performing functions, little room for improvement exists given the currently available data. Indeed, these functions fall into two categories, ribosomal processing and basic cellular metabolism, both of which are known to have clear 'global' signals in microarray data (Eisen *et al*., 1998; Jansen *et al*., 2002). That is, given a collection of microarray experiments performed under almost any conditions, it is likely that genes related to ribosomes and cellular respiration will be coexpressed at detectable levels. This ubiquitousness makes these functions easy to detect by techniques such as concatenation; even a modest signal present in most microarrays will be detectable in a correlation calculated across all experiments simultaneously. This accounts for the ease with which ribosomal function can be predicted from coexpression data (Gasch *et al*., 2000; Karaoz *et al*., 2004; Lanckriet *et al*., 2004). Other functions in which MEFIT shows little improvement, (such as *protein kinase cascade* or *fermentation*) are small or poorly studied functions in which data sparsity makes it impossible for any prediction technique to perform well.

Conversely, the functions in which MEFIT provides the most improvement tend to be specific functions that are reasonably well represented in the data but are poorly predicted by more global methods. For example, genes involved in the *meiotic cell cycle/response to pheromone/sporulation* group of functions should be coexpressed only under very specific circumstances; such a signal would be undetectable by correlation across all datasets simultaneously. Relative to CONCATENATION and Z-SCORE integration, the LLS method also provides some improvement for several such specific functions. Since MEFIT is designed to upweight datasets within functions where they demonstrate predictive power, this method is able to extract more localized signals originating in a few microarrays performed under appropriate conditions.

3.2 Bayesian integration of microarray datasets

It is of interest to compare the performance of the four global microarray integration methods—BAYESIAN, CONCATENATION, Z-SCORE and LLS—on the entire answer set, without decomposing the results into specific biological functions. Additionally,

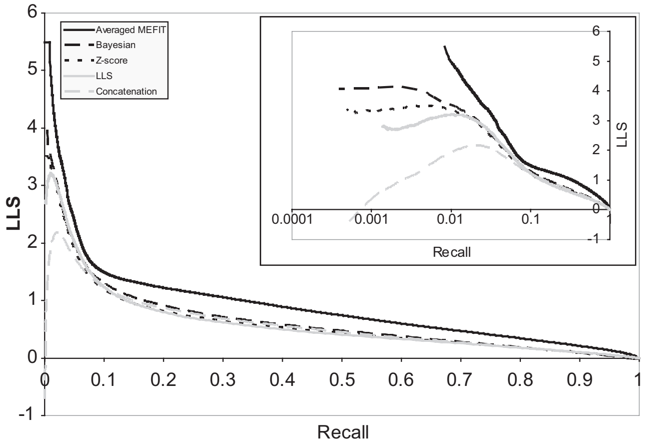

Fig. 3. Global integration performance. A comparison of the five global integration methods. A log scale inset is shown to emphasize the high precision area of biological interest; the minimum recall is limited to a minimum of 100 positive predictions to avoid noise. Performance is shown using the LLS score $LLS = \log_2(TP{\cdot}N/FP/P)$ for P total positive pairs, N total negative pairs and TP and FP the number of true and false positives at a particular sensitivity threshold.

one can reintegrate the individual components of the MEFIT output in a variety of ways to produce a global prediction set; relative to the BAYESIAN integration method, this has the benefit of preserving dataset/function associations by weighing each dataset by its relevance to each function. Ideally, each gene pair should be globally related if it is related in at least one function. In practice, noise in the predictions can make this assumption error-prone (a single overconfident prediction can dominate the overall probability), so we reintegrate each gene pair by taking the average probability of functional relationship over all functions in which that pair is predicted. This reintegration of the MEFIT output will be referred to as the AVERAGED MEFIT method.

Perhaps the most striking feature of a comparison of these global integration techniques (Fig. 3) is the sharp decline in precision of the CONCATENATION method at low recall (i.e. high correlation). In other words, gene pairs strongly correlated across an extremely large vector of disparate microarray conditions tend to be functionally unrelated. This is caused by factors such as transposable elements and similar sources of homology (telomeric sequences, etc.) that can lead to non-functionally related coexpression under essentially any experimental conditions. These sequences are known to be problematic due to cross-hybridization in coexpression experiments (Bozdech *et al*., 2003), and they are excluded from most microarrays; the CONCATENATION method interprets this absence as missing data and sees these genes only as very strongly correlated across the few datasets in which they are present, resulting in its poor performance at low recall.

Both the BAYESIAN and AVERAGED MEFIT results retain high precision at low recall cutoffs, with the AVERAGED MEFIT method also showing a substantial improvement in high recall areas. These integration techniques both explicitly encode the necessity to ignore or downweight inputs that tend to be overconfident (e.g. datasets in which a high correlation is not necessarily indicative of functional relationship), leading to their improved low recall behavior. AVERAGED MEFIT integration is able to perform the same downweighting on a per-function basis, which likely contributes to its greater precision at high recall.

3.3 Functional analysis reveals both expected and novel data content

In addition to per-function and global predictions of gene pair functional relationships, MEFIT also provides information on the relationships between microarray datasets and biological processes. Specifically, each per-function network learns during training how reliable it expects each dataset to be within its function. These reliabilities can be extracted as posterior probabilities after Bayesian inference, leading to a single confidence score for each dataset in each biological function.

Several aspects of these confidence scores (Fig. 4) demonstrate clear agreement with the per-function results shown earlier and with existing biological knowledge. Nearly every dataset for example is highly informative regarding *ribosome biogenesis* and *rRNA metabolism* (Fig. 4, light gray cluster), for the reasons discussed above; this is accompanied by a similar, weaker signal from the general *translation* function. Of the datasets in which ribosomal functions are not well predicted, (Angus-Hill *et al*., 2001) and (Rudra *et al*., 2005) are knockouts in which ribosomal genes are specifically disrupted.

Many datasets have moderately strong responses in *amine*, *amino acid* and *organic acid metabolism* (Fig. 4, black cluster), but the (Brem and Kruglyak, 2005) and (Yvert *et al*., 2003) results particularly stand out. These are both recombination studies between lab (BY4716) and wine (RM11-1a) strains with a focus on regulatory relationships. Other strong signals arise from the (Hardwick *et al*., 1999) study investigating rapamycin treatment and nutrient response and from the two (Saldanha *et al*., 2004) datasets for leucine and uracil limitation. All of these experiments have clear ties to amine and amino acid metabolism.

Three datasets are found to be particularly informative for *DNA recombination*, *M phase*, *meiotic cell cycle*, and *sporulation* (Fig. 4, dark gray cluster); these are (Primig *et al*., 2000) (a sporulation time course), (Williams *et al*., 2002) (UME6 deletion, a known meiotic regulator) and (Jin *et al*., 2004). While the first two findings seem logical, (Jin *et al*., 2004) studies xylose metabolism and fermentation, which has no obvious connection to meiosis. Our functional relevance results in this case alert a biologist to the possibility of a sporulation response to an inhospitable medium, a pre-sporulation response (Pringle *et al*., 1997) or a disruption of the nutrient response pathways due to introduction of the XYL1, XYL2 and XYL3 genes (Jin *et al*., 2004), information that might not have been evident without such a per-function analysis.

A biologist could use such information in at least two ways. Given a set of existing microarrays and a pathway or process of interest, this functional decomposition reveals datasets with an increased likelihood of containing information regarding that pathway or process. Conversely, given a new microarray (possibly generated under experimental conditions spanning many functions), functional decomposition produces a summary of pathways potentially disrupted or activated under its conditions. These analysis methods would be lost in an integration technique not taking advantage of the functional specificity of microarray datasets and of functional relationships.

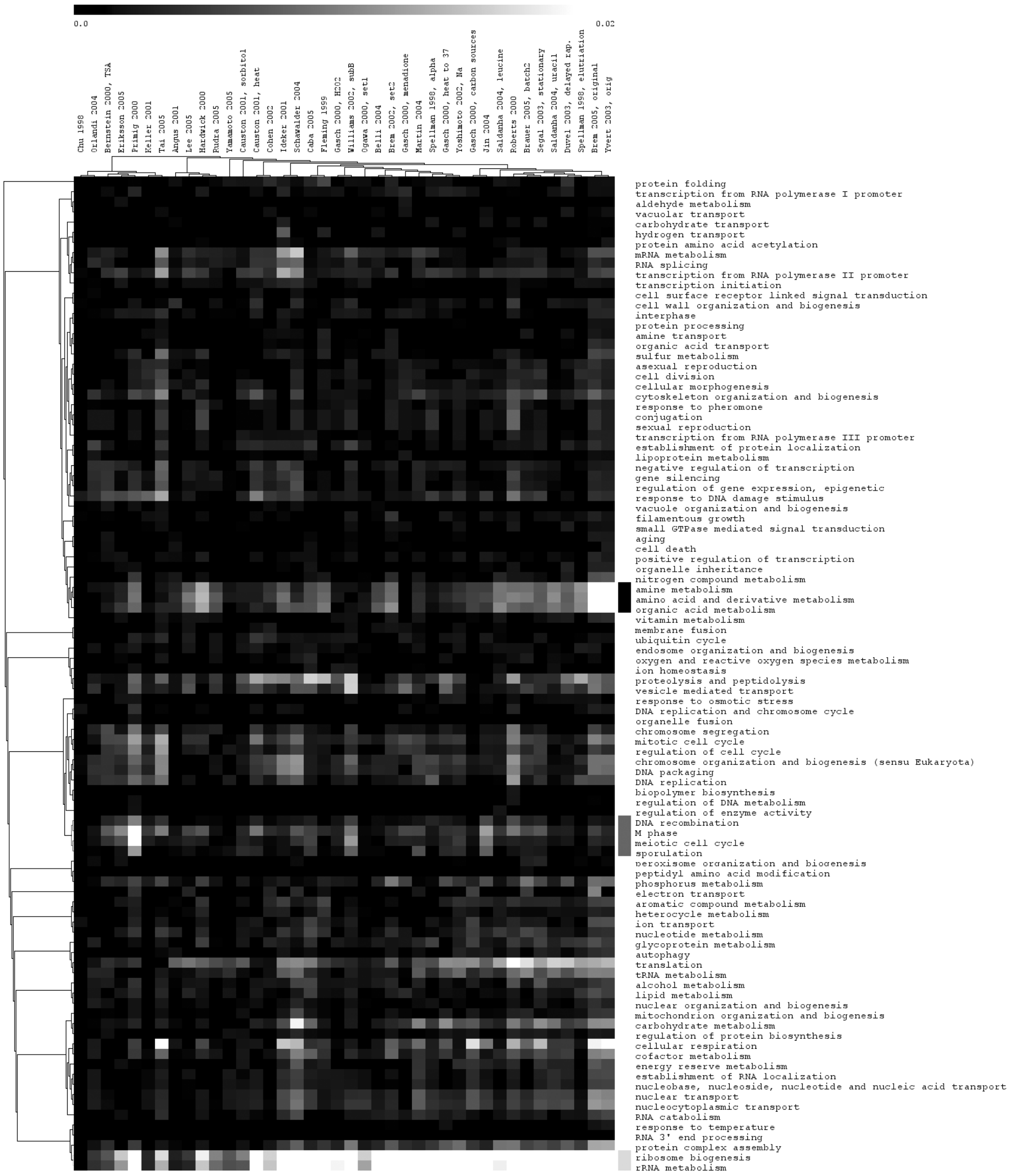

Fig. 4. Predictive power of datasets within individual biological functions A portion of the per-function dataset confidence scores learned by MEFIT. Brighter cells indicate a higher average posterior probability of functional relationship given input from a particular microarray dataset in a particular biological function. These are calculated from networks averaged over a 5-fold cross validation and are small due to the volume of microarray data employed (the maximum average difference for permuted data are ~0.005). Datasets and ontology terms have been clustered to visually show similarities in predictive power. The three colored clusters (amine metabolism in black, meiosis in dark gray, and ribosomal in light gray) represent interesting predictions discussed in the text. The heat map was generating using TIGR MeV (Saeed *et al.*, 2003).

3.4 Novel functional predictions

Based on the integrated per-function coexpression networks predicted by MEFIT, we can make functional predictions for genes previously unannotated in GO. Specifically, we examined several functions in which MEFIT showed marked improvement over previous integration techniques and extracted the most confident predictions. Searching these predictions for highly connected subgraphs involving both known and unknown genes produced several candidates, of which we chose to examine two: YML037C and YHR159W clustered around MMS4 in the *meiotic cell cycle* function, and YKR016W, YNL100W and YNL274C clustered around INH1 and TIM11 in *hydrogen transport*.

All six of these genes are uncharacterized open reading frames annotated to *biological process unknown*, save for YNL274C's overly general *metabolism* annotation (which was unused in our analysis). In *hydrogen transport*, the GO term representing mitochondrial proton processing, INH1 and TIM11 are both proteins associated with the F1F0-ATP synthase (Brunner *et al*., 2002). Of our predictions, YNL100W and YKR016W are known to localize to the mitochondrion (Huh *et al*., 2003), and all three appear in the mitochondrial proteome (Sickmann *et al*., 2003). Deletion of YKR016W also shows growth defects on non-fermentable carbon sources (Steinmetz *et al*., 2002), which we have confirmed in our lab (data not shown). YNL274C shows no strong localization, but its sequence contains a hydroxyacid dehydrogenase domain targeting NAD (Mulder *et al*., 2005), supporting our predicted role in cellular respiration.

Our predicted meiotic cell cycle group is centered on MMS4, which is a meiotic and mitotic gene involved in recombination and DNA repair (Xiao *et al*., 1998). YML037C shows a strong colocalization with clathrin coated vesicles (Huh *et al*., 2003), appears to behave as a transcriptional activator (Titz *et al*., 2006), and may be a substrate of the DBF2-MOB1 mitotic exit regulation complex (Mah *et al*., 2005). These characteristics point towards a potential mitotic or meiotic regulatory role for YML037C, in agreement with our prediction. YHR159W is thought to be a phosphorylation target of CDK1/CDC28, showing cell cycle regulation peaking in G1 (Ubersax *et al*., 2003); tests in our lab have shown that a heterozygous deletion mutant appears to be defective in tetrad formation during sporulation, a phenotype that we are currently investigating in more detail.

4 DISCUSSION

Here, we present MEFIT, a methodology for the simultaneous analysis of multiple microarray datasets using Bayesian integration augmented by per-function analysis. MEFIT's integration improves upon the general predictive power of existing methods for discovery of pairwise functional relationships from diverse microarray data. Additionally, it produces a per-function analysis for biologists, providing predictions in the context of individual pathways or biological processes (which may also be specified initially by the biologist).

Two strengths of MEFIT lie in its scalability and interpretability. Naive Bayesian learning and inference are both computationally inexpensive, and analysis can be performed simultaneously for hundreds of datasets spanning thousands of conditions. Additionally, given a single new dataset to integrate, no retraining need be performed—the conditional probabilities relevant to the new data can be learned independently of existing data. The statistics required for dataset normalization are fairly standard, and learned network parameters are readily interpretable and visualizable as probability distributions over each dataset and function.

As defined above, 'functions' in this framework are simply gene lists defined by some prior method to be functionally related. These might consist of pathways or transcription factor modules specified by a biologist or of larger collections of genes; we have used groups of genes sharing annotations in GO (as well as performing initial validations with the MIPS hierarchy). This could easily be extended into other organisms, for example by using tissue types or cancer pathways in mammalian systems. MEFIT learns to predict novel functional relationships similar to those specified in its input sets.

The output of MEFIT is one naive Bayesian network per-function; dataset to function confidence values and per-function probabilities of gene pair functional relationships can be immediately derived from these learned networks. In other words, MEFIT produces one genetic interaction network per function in addition to a global interaction network; if desired, by interpreting pairwise probabilities as similarity scores, these predictions can be further visualized (e.g. as per-function clusters). Since functional relationships are frequently specific to individual biological processes [such as STE7/FUS3 interaction during pheromone response versus STE7/KSS1 interaction during nutrient limitation (Madhani and Fink, 1997; Ptashne and Gann, 2003)], this provides a biological perspective that is both more realistic and, by compartmentalizing interactions, more manageable.

We have made our test predictions available at the MEFIT web site (http://function.princeton.edu/mefit/) along with a collection of predictions for the entire *S.cerevisiae* genome constructed by training on all known data and evaluating all gene pairs (including unknowns). This site includes an interface for browsing these predictions and the large collection of microarray datasets used to generate them. We expect that this microarray integration methodology will be useful in the context of heterogeneous data integration tools, where it can provide more informative preprocessing of coexpression data. We have already established substantial biological evidence for several of MEFIT's predictions, and we hope that it will provide a useful tool for guiding future laboratory and high-throughput experiments.

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Angus-Hill,M.L. *et al*. (2001) A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control. *Mol. Cell*, **7**, 741–751.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Baggerly,K.A. *et al*. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8**, 639–659.

Ball,C.A. *et al*. (2005) The stanford microarray database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.*, **33**, D580–D582.

Barrett,T. *et al*. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.*, **33**, D562–D566.

Bergmann,S. *et al*. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.*, **67**, 031902.

Bozdech,Z. *et al*. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol*., **4**, R9.

Brem,R.B. and Kruglyak,L. (2005) The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 1572–1577.

Brunner,S. *et al*. (2002) Su e of the yeast F1Fo-ATP synthase forms homodimers. *J. Biol. Chem*., **277**, 48484–48489.

Butte,A.J. and Kohane,I.S. (2006) Creation and implications of a phenome-genome network. *Nat. Biotechnol*., **24**, 55–62.

Butte,A.J. *et al*. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA*, **97**, 12182–12186.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 93–103.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.

Clare,A. and King,R.D. (2003) Predicting gene function in *Saccharomyces cerevisiae*. *Bioinformatics*, **19**, II42–II49.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*., **4**, 210.

David,F.N. (1949) The moments of the Z and F distributions. *Biometrika*, **36**, 394–403.

Druzdzel,M. SMILE: structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. *Proceedings of the Sixteenth National Conference on Artificial Intelligence* pp. 902–903.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Griffith,O.L. *et al*. (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics*, **86**, 476–488.

Griffith,J.S. *et al*. (1999) Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proc. Natl Acad. Sci. USA*, **96**, 14866–14870.

Heyer,L.J. *et al*. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*., **9**, 1106–1115.

Hu,P. *et al*. (2005) Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, **6**, 128.

Huh,W.K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Huttenhower,C. and Troyanskaya,O. (2006) Bayesian data integration: a functional perspective. *Comput. Syst. Bioinformatics*, in press.

Ideker,T.E. *et al*. (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput*., 305–316.

Ihmels,J. *et al*. (2005) Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet*., **1**, e39.

Ihmels,J. *et al*. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet*., **31**, 370–377.

Jansen,R. *et al*. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res*., **12**, 37–46.

Jin,Y.S. *et al*. (2004) *Saccharomyces cerevisiae* engineered for xylose metabolism exhibits a respiratory response. *Appl. Environ. Microbiol*., **70**, 6816–6825.

Karaoz,U. *et al*. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA*, **101**, 2888–2893.

Kloster,M. *et al*. (2005) Finding regulatory modules through large-scale gene-expression data analysis. *Bioinformatics*, **21**, 1172–1179.

Lanckriet,G.R. *et al*. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput*., 300–311.

Lauritzen,S. and Spiegelhalter,D. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc*., **50**.

Lee,I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Madhani,H.D. and Fink,G.R. (1997) Combinatorial control required for the specificity of yeast MAPK signaling. *Science*, **275**, 1314–1317.

Mah,A.S. *et al*. (2005) Substrate specificity analysis of protein kinase complex Dbf2-Mob1 by peptide library and proteome array screening. *BMC Biochem*., **6**, 22.

Moreau,Y. *et al*. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*., **19**, 570–577.

Mulder,N.J. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.

Mulligan,M.K. *et al*. (2006) Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. *Proc. Natl Acad. Sci. USA*, **103**, 6368–6373.

Neapolitan,R. (2004) *Learning Bayesian Networks*. Prentice Hall, Chicago, IL.

Pavlidis,P. *et al*. (2002) Learning gene functional classifications from multiple data types. *J. Comput. Biol*., **9**, 401–411.

Primig,M. *et al*. (2000) The core meiotic transcriptome in budding yeasts. *Nat. Genet*., **26**, 415–423.

Pringle,J. *et al*. (1997) *Saccharomyces:* cell cycle and cell biology. *Mol. Cell. Biol. Yeast*, in press.

Ptashne,M. and Gann,A. (2003) Signal transduction. *Imposing specificity on kinases, Science*, **299**, 1025–1027.

Rhodes,D.R. *et al*. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.

Rudra,D. *et al*. (2005) Central role of Ifh1p-Fhl1p interaction in the synthesis of yeast ribosomal proteins. *EMBO J*., **24**, 533–542.

Ruepp,A. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*., **32**, 5539–5545.

Saeed,A.I. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Saldanha,A.J. *et al*. (2004) Nutritional homeostasis in batch and steady-state culture of yeast. *Mol. Biol. Cell*, **15**, 4089–4104.

Sickmann,A. *et al*. (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl Acad. Sci. USA*, **100**, 13207–13212.

Steinmetz,L.M. *et al*. (2002) Systematic screen for human disease genes in yeast. *Nat. Genet*., **31**, 400–404.

Titz,B. *et al*. (2006) Transcriptional activators in yeast. *Nucleic Acids Res*., **34**, 955–967.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Troyanskaya,O.G. *et al*. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA*, **100**, 8348–8353.

Ubersax,J.A. *et al*. (2003) Targets of the cyclin-dependent kinase Cdk1. *Nature*, **425**, 859–864.

Williams,R.M. *et al*. (2002) The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast. *Proc. Natl Acad. Sci. USA*, **99**, 13431–13436.

Xiao,W. *et al*. (1998) Mms4, a putative transcriptional (co)activator, protects *Saccharomyces cerevisiae* cells from endogenous and environmental DNA damage. *Mol. Gen. Genet*., **257**, 614–623.

Yvert,G. *et al*. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet*., **35**, 57–64.